Behind Points and Levels — The Influence of Gamification Algorithms on Requirements Prioritization

Huber Kolpondinos, Martina Z ; Glinz, Martin

Abstract: Prioritizing requirements is a crucial ingredient of successful Requirements Engineering (RE). The popular prioritization techniques assume that stakeholders are known and can be mandated to contribute to the prioritization process. This prerequisite no longer holds for many of today's systems where significant stakeholders (end-users, in particular) are outside organizational reach: they are neither known nor can they be identified among the members of the involved organizations. Classic techniques for involving these stakeholders such as polls or questionnaires are neither interactive nor collaborative, which is detrimental for prioritization. Social media enable collaborative prioritization, but fall short in motivating stakeholders outside organizational reach to participate voluntarily. In this light, we are developing the Garuso platform, which combines social media with gamification for motivating stakeholders. While first approaches to employing gamification in RE are promising, research is still in its infancy. Especially, little is known about the influence of the gamification algorithms controlling single game elements on the stakeholders' activities. In this paper we report on a field experiment in which we investigated this influence with Garuso. We found statistically significant differences between different algorithms controlling single game elements on the contributions of stakeholders to the prioritization of requirements.

DOI: https://doi.org/10.1109/RE.2017.59

Behind Points and Levels – The Influence of Gamification Algorithms on Requirements Prioritization

Martina Z. Huber Kolpondinos*,†, Martin Glinz*
*Department of Informatics, University of Zurich, Switzerland
{kolpondinos, glinz}@ifi.uzh.ch
†Empa, Swiss Federal Laboratories for Materials Science and Technology, St. Gallen, Switzerland

***Abstract*—Prioritizing requirements is a crucial ingredient of successful Requirements Engineering (RE). The popular prioritization techniques assume that stakeholders are known and can be mandated to contribute to the prioritization process. This prerequisite no longer holds for many of today's systems where significant stakeholders (end-users, in particular) are outside organizational reach: they are neither known nor can they be identified among the members of the involved organizations. Classic techniques for involving these stakeholders such as polls or questionnaires are neither interactive nor collaborative, which is detrimental for prioritization. Social media enable collaborative prioritization, but fall short in motivating stakeholders outside organizational reach to participate voluntarily. In this light, we are developing the *Garuso* platform, which combines social media with gamification for motivating stakeholders. While first approaches to employing gamification in RE are promising, research is still in its infancy. Especially, little is known about the influence of the gamification algorithms controlling single game elements on the stakeholders' activities. In this paper we report on a field experiment in which we investigated this influence with Garuso. We found statistically significant differences between different algorithms controlling single game elements on the contributions of stakeholders to the prioritization of requirements.**

***Index Terms*—gamification; requirements prioritization; stakeholder motivation; field experiment;**

I. Introduction

Successful development and deployment of a software system crucially depends on knowing the priority of the elicited requirements [1]. Involving the stakeholders [2] of the system in the prioritizing process, in particular the end-users, increases the success of the software system [3]. As stakeholders typically do not have the same needs, they should be able to contribute collaboratively to the prioritization process [4], [5].

In requirements engineering (RE) research and practice, a wealth of prioritization techniques [1] have been developed and applied. The most popular and successful ones, such as ranking or grouping assume that the stakeholders are known and available. Hence, with the transition from dedicated software systems used by trained users to today's world of ubiquitous apps and globally offered services, the established prioritization techniques are seriously challenged. Significant stakeholders of these systems, in particular, end-users, are typically not known and also outside organizational reach, i.e., they cannot be identified among the members of the involved organizations. If these stakeholders are ignored, valuable knowledge may be missed [3]. Online polls and questionnaires are established technical means for involving such stakeholders. As neither of them is collaborative, they are not well suited for requirements prioritization. More recently, social media have been proposed for performing collaborative RE tasks such as requirements prioritization [6], [7]. However, social media approaches do not addresses the motivation problem: stakeholders outside organizational reach need to be motivated to contribute voluntarily, as they cannot be mandated to contribute.

Here, gamification, the use of game elements in non-game contexts, can provide a solution. First approaches applying gamification in RE are promising, e.g., [8], [9]. However, this research is still in its infancy: little is known about the influence of single game elements and the algorithms controlling them on the stakeholders' RE activities, and nothing with respect to stakeholders outside organizational reach. This may lead to mistakes when applying gamification, which bears the risk of damaging the stakeholders' inherent motivation [10].

In this context, we are developing the Garuso (**G**ame-b**a**sed **R**equ**i**rement**s** Elicitati**o**n) platform, which combines a forum for contributing, discussing and rating needs with game-based techniques for motivating potential stakeholders to contribute to these RE activities.

In this paper, we report on the results of a field experiment in which we investigated the influence of the algorithms that control the popular game elements *points* and *levels* on the contributions of stakeholders outside organizational reach to the prioritization of requirements. The experiment was conducted on the Garuso platform. We found that using different algorithms indeed has a statistically significant influence on the contributions of stakeholders to prioritization.

The remainder of this paper is structured as follows. We provide background information and related work in Sect. II. In Sect. III, we give an overview of Garuso. Then we describe the experiment in Sect. IV. The results are presented and discussed in Sect. V-VIII. Sect. IX concludes the paper.

II. Background and Related Work

This section provides background information and related work on prioritization and gamification in RE. Further, we motivate the need for studying the gamification algorithms.

Authors' preprint version
25th IEEE International Requirements Engineering Conference (RE'17). pp. 332–341.

A. Requirements Prioritization

Prioritizing requirements means to determine their relative necessity [1] with respect to business goals, available resources, and existing constraints [11]. It is an iterative process that can be performed during the entire lifecycle of a software system [12]. The prioritization techniques used by requirements engineers and practitioners are well established and manifold [13], [1]. For example, requirements can be ranked by multiple criteria, e.g., Cost-Value Ranking [14], or in relation to other requirements, e.g., Pairwise Comparison [15]. For achieving scalability, these techniques can be supported with data mining and machine learning techniques [13].

Recent approaches in requirements prioritization increasingly focus on social interactions that collaboratively involve all stakeholders [16]. For example, WikiWinWin [4] enables quick collaboration on the Web. Stakeholders can brainstorm new needs collaboratively and rate each others' contributions with respect to different predefined criteria, e.g., business importance or ease of realization. Online platforms typically enable the stakeholders to prioritize contributed needs in a more sophisticated way. For example, the approach by Lohmann et al. [6] enables stakeholders to rate shared needs on a scale and also to vote for or against them. Similarly, the collaborative RE framework by Konaté et al. [5] uses two consecutive prioritization steps: (1) voting for or against needs based on a personal perception of importance; (2) selecting key needs among those that received the most votes. Most recently, Liquid RE [7] suggests to grant the stakeholders the right to delegate their vote to other stakeholders.

However, motivating stakeholders, particularly those outside organizational reach, towards voluntarily contributing to requirements prioritization is still an open issue. Here, gamification offers an interesting chance.

B. Gamification in Requirements Engineering

Gamification is the use of game elements in a non-game context [17]. It harnesses the motivational power of games and applies it to real-world problems [18]. A crucial prerequisite for the success of gamification is that people already have an inherent motivation towards the product or service to which gamification is applied [19]. Stakeholders have by definition an interest in the software system under consideration [2] and therefore meet this prerequisite.

The involvement of end-users in RE activities has been identified as a key challenge for the success of a software system [20]. Recently, requirements engineers have started to address this challenge with gamification. First approaches that apply gamification in the context of requirements elicitation and prioritization show encouraging results with respect to the engagement of stakeholders *within* organizational reach. For example, two case studies involving the web-based gamification environment iThink [8] yielded highly satisfying results with regard to the number and quality of the generated requirements. Similarly, results of a more recent case study involving the online platform REfine [9] show a positive influence of gamification on collaborative RE activities such as suggesting, branching, and prioritizing needs and comments. Most recently, in the context of scenario-based RE, the results of a controlled laboratory experiment showed that the participants who were motivated with game elements on a digital platform produced user stories that led to requirements of higher quality and creativity than those produced without gamification [21].

However, to the best of our knowledge, no studies on the involvement of stakeholders *outside* organizational have been published so far.

C. The Need for Investigating Single Game Elements

To be motivated is a delicate state on a scale between no motivation (amotivation) and absolute motivation (inherent motivation) [22]. Badly designed motivation strategies can push inherently motivated people towards the state of amotivation, e.g., by overjustification [23]. Examples of badly designed gamification include designs that control the users too much, i.e., give them no autonomy on their activities, or provide rewards that are meaningless for them. In particular, the random application of game elements has been criticized for achieving results below expectations and for dulling the users [10].

In software engineering (SE), the lack of a systematic methodology on how to apply gamification to increase user engagement has been identified as a research gap and threat [24]. Also, the small number of studies researching the effects of single game elements (compared to the number of studies on general effects of gamification, i.e., regarding all applied game elements together as one black box) has been criticized [25]. In RE, the need to investigate the effects of gamification on stakeholder engagement more thoroghly also has been recognized. For example, the creators of the iThink approach [8] (see above) identified the lack of an experiment as a limitation of their work [26]. Further, researchers who investigated general effects of gamification have emphasized the need for testing game elements in isolation [21].

III. Garuso

The experiment we report on in this paper was conducted on the Garuso platform that we are developing at the University of Zurich. To understand the context of the experiment, we briefly describe the architecture of the Garuso platform (Fig. 1), its user interface (Fig. 2), and the rating scheme used for prioritization (Fig. 3).

Garuso (**G**ame-b**a**sed **R**equ**i**rement**s** Elicitati**o**n) is a research project that investigates stakeholder engagement with respect to the collaborative elicitation and prioritization of requirements. The conceptual basis of Garuso is a three-dimensional motivation concept [27] that we created based on theories of experiential learning and motivational psychology.

A. The RE Module

The *RE module* addresses asynchronous communication and creative contributions. It offers four RE related features that facilitate the collaborative elicitation and prioritization of

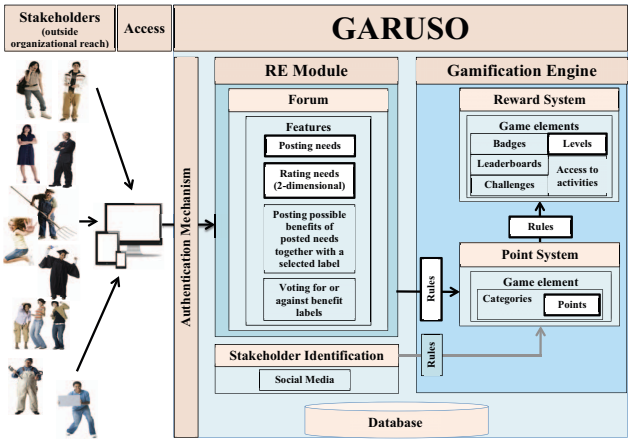

Fig. 1. The architecture of the Garuso platform. The features in bold frames indicate the activities and game elements that were enabled in the experiment.

requirements with respect to the software system of interest: on the Garuso platform, stakeholders can (1) post their needs with text and/or images, (2) rate each others' needs, (3) post and label benefits for all the posted needs, and (4) vote for or against the labels of these benefits. For the experiment, we limited the available features to (1) and (2).

B. The Gamification Engine

The *gamification engine* offers means to motivate stakeholders towards using the features offered by the *RE module*. It includes a point system, a reward system, and rules.

1) Points and Rewards: The *point system* uses the game element *point* and defines different point categories. It is directly affected by the activities performed by the platform users. The *reward system* uses the game elements *badges*, *leaderboards*, *challenges*, *levels*, and *access* to reward the users. It is built on top of the point system and directly affected by a user's earned number of points per category. For the experiment, we limited the game elements to points and levels.

2) The Rules: The *rules* govern the game elements in the point system and the reward system. They define how many points of which categories users *earn* for their activities on the platform, and how many points of which categories they *need* for each reward. The rules are implemented with *algorithms*.

C. Stakeholder Identification

The *stakeholder identification* module enables stakeholders to invite other potential stakeholders over different social media channels to participate. This approach is known as snowballing [28] and was previously applied in RE for stakeholder identification [29]. To ensure equal basic knowledge of the participants, we did not enable stakeholder identification for the experiment.

D. The Garuso User Interface

Fig. 2 shows a screenshot of the user interface (UI) main page for user *Feta*, one of our experiment participants. The left sidebar shows the user engagement: earned points are displayed in the upper part and the percentage of rated posts in the lower part. *Feta* has currently earned 1000 points and rated 95 percent of all posts. The center part shows a welcome message, a button for creating posts, and a search field, followed by all posts of Feta's group in pages of nine posts each. To balance their visibility, the posts are randomly ordered over time. Further, the pages can be switched manually. Posts that the user has not yet rated are displayed in rose, while the rated ones are displayed in green. The user's own posts (which she or he cannot rate) are colored in blue. In Fig. 2, *Feta* has rated five posts and one post left to rate. The right sidebar shows the user competence. The user's current level is displayed graphically in the upper part and her or his status in the lower part. User *Feta* is currently on level one, needs four more rating values for achieving level two, and may lose 23 rating values until falling back to level zero.

E. Rating Scheme and Rating Values

The rating of posts is facilitated with a two-dimensional rating scheme (Fig. 3). The rating scheme is available in the detailed view of a post, which opens when clicking on a post on the UI main page. It offers ten rating options that each are represented by a button. Nine options are grouped in a matrix where the x-axis denotes the popularity and the y-axis the relevance of the post as perceived by the user. For example, the top right button indicates that the post is liked and relevant to the software system of interest. The tenth option allows users to express that they do not want to rate a post.

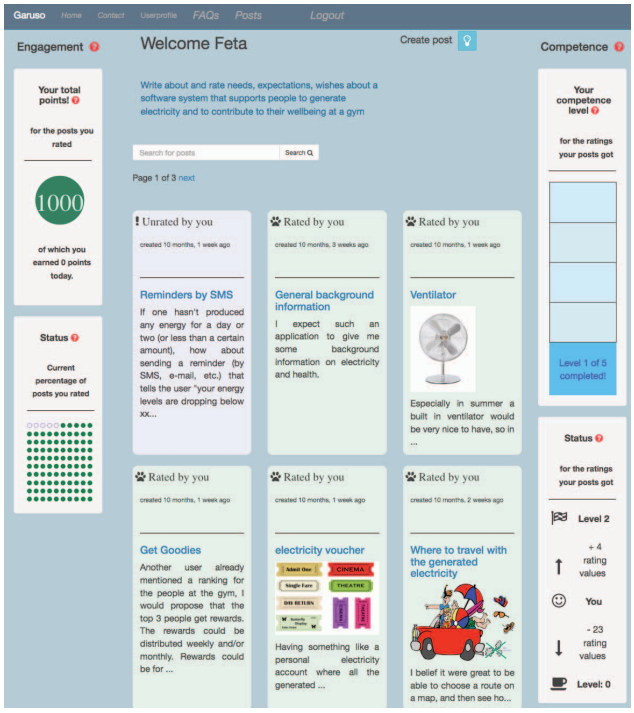

Fig. 2. The UI of the Garuso platform. Left-sidebar: point overview; center: posts, search field, button to create a post; right-sidebar: level overview

Fig. 3. The rating scheme with nine buttons to rate the relevance and the popularity of a post and one button to confirm not wanting to rate the post.

Users can change their ratings at any time. With this feature, we take the natural flow of interaction into account, which can also be observed in group elicitation methods [30] with physically present participants.

The rating values depend on the selection in the rating scheme. Both dimensions of the matrix in that scheme have a value range of [-1,1]. We take the sum of both dimensions, yielding a range of [-2,2] for the rating value. For example, the rating *irrelevant&dislike* yields a rating value of minus two, *relevant&like* yields two, *relevant&neutral* yields one, and a *neutral&neutral* rating yields zero. *No rating* also yields a value of zero.

IV. The Experiment

We ran a field experiment on the Garuso platform to study the influence of the algorithms controlling the individual game elements of the *Gamification Engine* on collaborative requirements prioritization, which ist a typical RE activity performed on the platform. As game elements, we chose *points* and *levels* which belong to the most popular elements in gamification [17], [31] and have a high probability of also being known to people outside of a game context. The platform features required for prioritization are *post* and *rate*. With respect to the nature of an experiment, in particular for isolating the dependent variables, we disabled all other features of the platform.

The software system for which the participants posted and rated needs on the Garuso platform, is part of a smart living project [32] in which the energy produced by people when working out in a gym is used to generate electricity. The purpose of the software system is to motivate people towards using such enhanced workout equipment. We call it the *Smart Workout Motivation System* or *SmaWoMo* for short in the remainder of this paper.

Twenty people participated in the experiment that we ran over a period of twelve days from July to August 2016.

A. Goal, Research Questions and Hypotheses

We followed the Goal-Question-Metric (GQM) approach [33] to define our goal and refine it into research questions and hypotheses as presented in the following. The metrics we used are explained in sub-section IV-D.

Our goal is to ***investigate*** *the effects that algorithms controlling game elements have on RE activities undertaken by users of the Garuso platform* ***for the purpose*** *of collaborative requirements prioritization* ***with respect to*** *user engagement and user acceptance* ***from the point of view*** *of stakeholders outside organizational reach* ***in the context of*** *a field experiment.*

We address this goal with the following two research questions and the corresponding hypotheses.

RQ1: What is the influence of algorithms calculating the number of points stakeholders get for rating needs posted by others on the engagement of these stakeholders?

H_0-1: Algorithms calculating the points stakeholders get for rating others' posted needs have no influence on the engagement of these stakeholders.

RQ2: How do algorithms mapping the values of the ratings the stakeholders received for their posted needs to levels influence the average stakeholder acceptance of posted needs?

H_0-2: Algorithms mapping the sum of received rating values of posted needs to levels have no influence on the average stakeholder acceptance of posted needs.

B. Experiment Design

To test our hypotheses, the first author of this paper conducted a field experiment with a between subject design [34]. Therefore, the participants were randomly assigned to the treatment group (TG) or control group (CG). The two groups used different instances of the Garuso platform, i.e., they did not interact with each other in any conceivable way, and we did not tell them about the existence of two groups. Further, all contributions were anonymous, e.g., the author and rating values of others' posts were not disclosed. Initially, we seeded three equal posts in both groups and repeated this step three times during the experiment to ensure the participants had enough posts to rate.

To properly define the variables and metrics we first discussed the experiment design within a group of senior researchers in the fields of RE, HCI, and Psychology, and implemented the algorithms and layout of the Garuso platform based on these results. Second, we informally tested the usability of the Garuso platform by involving these senior researchers.

To strengthen the conclusion that can be drawn from the experiment, we split the field experiment for each of the groups in two sub-experiments. The *point sub-experiment* considered the direct aspect of rating, i.e., the number of points a participant earns when rating a post for the first time. The *level sub-experiment* considered the indirect aspect of rating, i.e., the value the author of the post earns based on the rating choice made by the participant who rates the post.

C. The Participants

We recruited the participants from a group of 120 people who had participated in a previous online survey [35] about the Smart Workout Motivation (SmaWoMo) system and had indicated their interest in a follow-up activity. We sent an e-mail message to these people, informing them that they could further express and discuss their needs about the SmaWoMo system on the Garuso platform during a period of twelve days. The message included a link to the registration page of the Garuso platform. 23 persons actually registered. However, two

of them did not contribute anything and one only registered when the experiment was already over. So we had 20 people who actively participated in the experiment.

Due to the selection process, the participants can be considered to have the same basic knowledge about the SmaWoMo system. At the beginning of the experiment we only told them that their task was to discuss their needs with respect to SmaWoMo by contributing and rating posts on the Garuso platform. We did not disclose the existence of an experiment and never made any suggestions to perform certain activities.

To reflect the real world situation of an arbitrary group of stakeholders outside organizational reach, the participants were randomly assigned to the treatment group (TG) or to the control group (CG) when registering. As the two non-contributing registrants had been assigned to the TG, we eventually had nine people in the TG and eleven in the CG.

At the first login, the participants were asked to complete a short questionnaire. All of them did so. The results are summarized in Table I and explained below. As mentioned above, all participants had participated in a previous online survey. For that survey, we had sought participants over multiple channels, in particular: (1) a mass e-mail via the distribution office of the University of Zurich, (2) a public Facebook post, and (3) the intranet of our research partner Empa. From the 20 participants in our current experiment, sixteen had been found initially over channel (1), two over channel (2) and two over channel (3). The majority of the participants were completely unknown to the authors of this paper, i.e., we did not have any known connection to them. Due to these characteristics, all participants can be considered to be stakeholders outside organizational reach. With respect to demographics, the two groups were pretty well balanced. Concerning application domain knowledge, the groups were overall balanced, with some differences in the sub-domains: four participants in the CG have never performed workouts, while all participants in the TG have workout experience. On the other hand, the number of participants perceiving themselves as knowledgeable about renewable energies is higher in the CG than in the TG.

TABLE I
OVERVIEW OF THE PARTICIPANTS

		CG	TG
	Number	11	9
	Completely unknown	7	4
Initial Contact Channel	Mass e-mail	8	8
	Facebook	1	1
	Intranet	2	0
Demographics	Countries[1]	3	4
	Average Age	32	31
	Gender (female/male)	6/5	4/5
Application Domain: *Performing workouts*	Never	4	0
	Not anymore	4	5
	Currently	3	4
Application Domain: *Knowledge about renewable energies*	Below average	2	3
	Average	7	4
	Above average	2	2
	Expert	0	0

[1]Participants per country: CG: CH:7, DE:3, GR:1; TG: CH:6, BG:1, IT:1, US:1
CG: Control Group, TG: Treatment Group

D. Variables and Metrics

To test our two hypotheses, we divided our experiment into two sub-experiments with an independent and a dependent variable each. The independent variables are the algorithms that control the game elements *points* and *levels*. The design of these algorithms follows the strategy of *experiencing success*, which is a common strategy for motivating players in game design [36] and users in gamification [37]. We considered two aspects of this strategy: (1) the aspect of *mastering challenges*, which is related to exploring; (2) the aspect of *fast progress*, which is related to achieving [38]. For both sub-experiments, the independent variables are summarized in Table II and subsequently explained together with the dependent variables.

1) Point Sub-Experiment: The independent variable in this sub-experiment is the algorithm that defines the number of points a participant earns for rating a post. We tested two values of this variable by using different algorithms for the TG and the CG. The TG algorithm uses a *binary function*, which addresses *mastering challenges*: per day, the number of points a participant receives is either zero as long as the participant has not rated all posts, or 100 as soon as the participant has rated all posts. The CG algorithm addresses *fast progress* with a *linear function*: per day, the number of points a participant receives is proportional to the percentage of the posts (s)he has rated. Both functions are normalized with the same maximum of points that can be earned per day. When the maximum is reached, it cannot be lost again.

The dependent variable in this sub-experiment is the *stakeholder engagement*. We argue that the number of posts a participant rates is an indicator for engagement. We measure this variable by calculating the number of all ratings as follows: (a) For visualizing participant behavior over time, we measure the average number of ratings per logged in participant for every day. (b) For hypothesis testing, we measure the total number of ratings for every participant over the full duration of the experiment.

2) Level Sub-Experiment: The independent variable in this sub-experiment is the algorithm that determines the competence level that a participant reaches based on the sum of all rating values that the posts of this participant have received from other participants. Again, we tested two values of this variable by using different algorithms for the TG and for the CG. Both algorithms address *mastering challenges* by initially

TABLE II
INDEPENDENT VARIABLES (ALGORITHMS) WITH RESPECT TO THE TWO SUB-EXPERIMENTS AND THE TWO EXPERIMENT GROUPS

	Control group	Treatment group
Point Sub-Experiment	Linear function: points equal to percentage of rated posts	Binary function: 0 points for rating $<$ 100% posts; 100 points for rating all posts
Level Sub-Experiment	Slowly increasing difficulty up to level four, decreasing difficulty for reaching level five	Rapidly increasing difficulty up to level three; decreasing difficulty above level three

increasing the difficulty to achieve the next level and *fast progress* by then switching to decreasing the difficulty for achieving the highest levels. The algorithms differ in the deltas required to achieve the next levels. This approach is also found in the literature, e.g., [36] and in existing systems, e.g. Stack Overflow [39]. In the TG algorithm, we increase the delta to reach the next level up to level three (26 rating values to reach level two, 36 rating values to reach level three) and then progress with decreasing deltas (29 and 7 rating values to reach levels four and five, respectively). In the CG algorithm, levels two and three are easier to achieve than in the TG (with deltas of 10 and 24 rating values), while achieving levels four and five is more difficult (with deltas of 36 and 28 rating values). For both groups, the calculation of the rating values is equal (cf. Sect. III-E) and the deltas are normalized for the levels one and five with two points and 100 points, respectively.

The dependent variable in this sub-experiment is the *stakeholder acceptance* of posts. We argue that the higher the value of a rating given by other participants, the higher is the acceptance of the post. We measure this variable by calculating the sum of all rating values as follows: (a) For visualizing participant behavior over time, we measure the average cumulative value per post and registered participant for every day. (b) For hypothesis testing, we measure the total value for every participant over the full duration of the experiment.

V. Data Collection and Analysis

During the experiment we monitored all user activities on the Garuso platform and stored the data in a database for subsequent analysis [40].

We analyzed the data in three ways: (1) we calculated the average of the metrics relevant to evaluate the dependent variables for both groups to investigate how the samples are represented (Table III); (2) we plotted the values of the dependent variables in both sub-experiments over the twelve days of the experiment to see how the values changed over time (Fig. 4 and 5); (3) for testing our hypotheses, we analyzed the values of the two dependent variables for every individual participant in the TG and in the CG es for both groups to investigate how the samples are represented (Table IV).

If a participant contributed continuously over the duration of the experiment, thereby producing a total of n ratings, we consider this to be a stronger engagement than that of a participant who logged in just a few times, also producing a total of n ratings. The same consideration applies for the total sum of rating values that a participant received. Thus, we normalized our data for each participant with the number of login days vs. total number of experiment days before we tested the hypotheses:

$$value_{normalized}(p_i) = value_{observed}(p_i) * \Sigma[login\ days]/[total\ days]$$

where p_i is the ith participant and $total\ days = 12$.

To determine a proper test for our hypotheses, we conducted a pre-evaluation in which we tested the data for normal distribution and equality of variances. The results are presented in the row labeled *Pre-Evaluation* of Table IV. Due to the small sample sizes we used the one-sample Kolmogorov-Smirnov (KS) test [41]. For both sub-experiments the KS test yielded a result with $p > 0.05$, i.e., not significant. Thus, we can assume normal distribution for all our data. The Levene test that we performed next yielded $p > 0.05$. Therefore, we can assume equality of the variances. Based on these results we ran the t-test on the hypotheses for both sub-experiments. To conclude the hypothesis testing, we evaluated the magnitude of the test results by calculating the effect size (the Pearson correlation coefficient) and classifying it according to Cohen [42].

VI. Results

The results of the two sub-experiments answer our research questions and give strong evidence that the way how algorithms are (reasonably) applied within a game-based elicitation platform has an influence on the contributions of stakeholders outside organizational reach to requirements prioritization. We first give some descriptive data for the two sub-experiments. Then we present the results of the two sub-experiments. Finally, we report on the results of a follow-up survey.

A. Overall Descriptive Data

In Table III we present the average values for login days, ratings, and posts as well as the average rating values for both the TG and the CG. These results indicate that the different gamification algorithms had an influence on the performance of the participants. While the rating values per post are similar for both groups, the number of activities per participant are higher within the TG. Subsequently, we present the detailed results for the two sub-experiments which confirm that the gamification algorithms indeed influence requirements prioritization.

B. Point Sub-Experiment

In the point sub-experiment we investigated the influence of gamification algorithms on the *stakeholder engagement* with respect to requirements prioritization by measuring the number of ratings that posts on the Garuso platform received. The results indicate that the way how gamification algorithms calculate the number of points earned for rating requirements has a significant influence on the number of ratings.

TABLE III
Average Data of the control and the treatment group

Metrics	Control Group	Treatment Group
#login days[1]	4.09	6.11
#ratings[1]	11.09	21.33
#posts[1]	1.27	2.22
Σ[rating values][2]	7.21	7.85

[1]Per participant [2]Per post

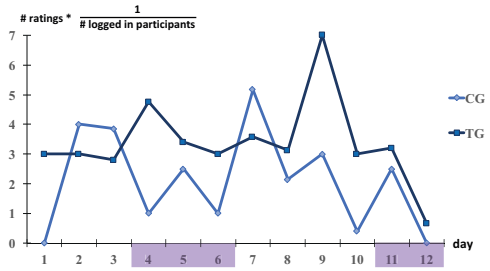

Fig. 4. Number of ratings per logged in participant for every day. The weekends and a public holiday (day 6) are marked with grey.

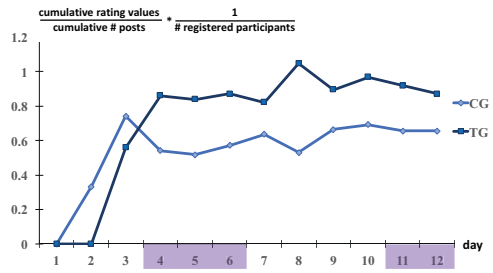

Fig. 5. Average cumulative rating values per post and participant for every day.

1) Observation: Figure 4 shows the number of ratings per day. We were interested in individualized results and therefore normalized this value for each day with the number of participants who had logged in on that day.

The graphs of the two groups have similar tendencies, but different characteristics. The values are high in the first two days, fluctuate within a range of approximately four between days four and nine, and significantly decrease afterwards. Both graphs increase three times in the second week and have their maximum peak in the second half of the experiment. The major difference appears between day three and four. Here, the sum of ratings in the TG increases while the corresponding value in the CG decreases and then remains lower than in the TG except for one day.

2) Hypothesis Testing: To investigate if the differences of the observed effects between the two groups are significant, we tested the corresponding hypothesis with a t-test.

H_0-1: Algorithms calculating the points stakeholders get for rating others' posted needs have no influence on the engagement of these stakeholders.

Table IV summarizes the test results. The descriptive statistics of the TG (μ=11, σ=6 with n=9) in which participants only earned the daily maximum of 100 points when rating all the posts of a day are higher compared to the ones of the CG (μ=5, σ=5.5 with n=11) in which participants earned points equal to the percentage of the posts they rated. The result of the t-test is significant at $p \leq 0.05$, so we can reject our null hypothesis. The effect size for this result is r=0.47, which represents a medium effect. This shows that the significance of our test results is meaningful.

C. Level Sub-Experiment

In the level sub-experiment, we studied the influence of gamification algorithms on the average *stakeholder acceptance* of posts by measuring the total rating value that the posts received on the Garuso platform. The results indicate that the way how gamification algorithms map the values of ratings for posted needs to levels has a significant influence on the acceptance of the posted needs by the stakeholders.

1) Observation: Figure 5 shows the average cumulative rating value per number of posts and day. We were interested in the cumulative results over a period of time and therefore normalized this value for each day with the number of participants who had registered up to that day. Again, the two graphs have similar tendencies, but different characteristics. They start with a steep slope and indicate that the average acceptance of a post per participant converges over time. Two major differences can be observed. First, after day three, the participants of the TG start to perform better compared to the ones of the CG, i.e., their posts were rated higher, and keep performing better until the end. Second, around day eight the values of the TG increase and the ones of the CG decrease.

2) Hypothesis Testing: To investigate the significance of the differences between the observed data of the two experiment groups, we tested the corresponding hypothesis with a t-test.

H_0-2: Algorithms mapping the sum of received rating values of posted needs to levels have no influence on the average stakeholder acceptance of posted needs.

The results are summarized in Table IV and further explained below. Regarding the average stakeholder acceptance of posts, the TG (μ=4.3, σ=1.9, n=9) performed better than

TABLE IV
OVERVIEW OF THE EXPERIMENT RESULTS

		Point Sub-Experiment		Level Sub-Experiment	
		Control Group	**Treatment Group**	**Control Group**	**Treatment Group**
Descriptive Statistics	μ	5.02	11	1.52	4.32
	σ	5.54	6.04	1.7	1.91
Pre-Evaluation	Normal Distribution[1]	0.139	0.2	0.2	0.2
	Variance equality[2]	0.796		0.589	
Significance Test[3]	p-Value	0.033		0.003	
	t-Value	-2.309		-3.469	
Magnitude	Effect Size[4]	0.478		0.633	

[1]Kolmogorov-Smirnov test (p-value) [2]Levene test (p-value) [3]t-test [4]Pearson (correlation coefficient r)

the CG (μ=1.5, σ=1.7, n=11). Recall that for the participants in the TG the difficulty to reach a competence level rapidly increased up to level three, while for the CG, the difficulty slowly increased up to the same maximum until level four. The result of the t-test is significant at $p \leq 0.05$, thus we can reject our null hypothesis. The effect size for this result is r=0.63 which represents a strong effect. This shows that the significance of our test results is meaningful.

D. Post-Experiment Survey

After the experiment, we sent an online questionnaire via e-mail to all participants asking them about their attitude towards the experiment and towards the influence of the game elements on their activities. Although they were offered an incentive, only 14 people, seven of each group, answered. The results are summarized in Fig. 6 and Fig. 7. To derive the participants' attitudes we followed the idea of semantic differential scales [43] with a one-polar scale, i.e., we used single adjectives instead of opposite pairs, where 1 means *not at all* and 7 means *absolutely*.

The results presented in Fig. 6 show similar perceptions in the CG and in the TG. The majority in both groups perceived the experiment as fairly interesting and fun, and as rather moderately exhausting and challenging.

The results shown in Fig. 7 are inconclusive. In both groups, more participants perceived the influence of points on the rating of posts to be stronger than the influence of the levels on the creation or style of posts. However, for both groups the answers are widely spread.

VII. Discussion

When revisiting our research questions, we can indeed state that the chosen algorithms do have an influence on the performance of the stakeholders.

A. Research Questions

RQ1: What is the influence of algorithms calculating the number of points stakeholders get for rating needs posted by others on the engagement of these stakeholders? The results of the *point sub-experiment* show that the way how algorithms calculate points for rating needs has a statistically significant effect on the prioritization process with respect to *stakeholder engagement*. The effect size indicates a medium effect of this result. Moreover, the results demonstrate that the influence was stronger in the treatment group where it was harder for the participants to earn the points than in the control group.

RQ2: How do algorithms mapping the values of the ratings the stakeholders received for their posted needs to levels influence the average stakeholder acceptance of posted needs? The results of the *level sub-experiment* show that the way how algorithms map rating values to levels has a statistically significant effect on the prioritization process with respect to the *stakeholder acceptance* of posted needs (according to the participants' perception). The effect size indicates a strong effect of this result. The results further demonstrate that the observed influence was stronger in the treatment group, where the difficulty to achieve a level rapidly increased until level three.

B. Overall Considerations

We found three overall aspects that we briefly discuss below. (1) On average, participants in the TG performed more activities than those in the CG. This result is surprising as the gamification algorithms that we used for the TG require more engagement at an early stage to reach the next goal, i.e., the next level, and the daily points, compared to the algorithms applied in the CG. A possible explanation for this behavior is that the participants of the TG were boosted by the higher challenge. Another explanation could be that the inherent motivation towards the Smart Workout Motivation (SwaWoMo) system was lower in the CG than the one in the TG. For example, four participants of the CG had never performed workouts, while all participants in the TG had. (2) In contrast to the different numbers of performed activities, we observed that the participants' behavior in performing these activities have similar tendencies in both groups over time. A possible reason for this result is that applying the same game element for the same task within the same context may lead to a similar user behavior (which is reflected by the similar graphs). (3) We cannot confirm an influence of weekends (days four to five and eleven to twelve) and holidays (day six) on the results. For days four to six, the values in the CG are below average, but the ones in the TG are not. On the other hand, the values decrease between days eleven and twelve in both groups. Yet, the latter effect could also be due to the end of the experiment.

The results of the experiment are not clearly supported by the results of the follow-up survey, where we found no major

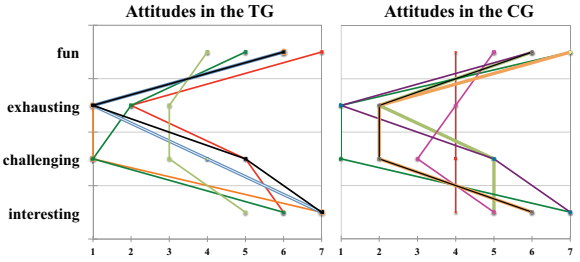

Fig. 6. Participants' attitudes towards the experiment.

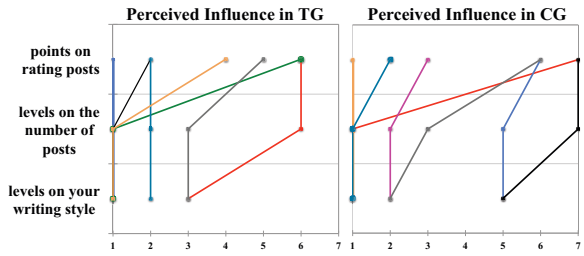

Fig. 7. Participants' perceived influence of game elements on activities.

difference with respect to the participants' attitude between the two groups. This contradicts the assumption that participants of the CG were less motivated to participate. Further, the survey results do not show any major difference between the two groups concerning the influence of the game elements. The CG members even perceived a slightly higher influence of the points received for rating than the TG members. This discrepancy between perception and reality might indicate that the participants were not aware of how much they were influenced by the game elements. Further, participation in the follow-up survey was not mandatory and only seven participants of each group completed the questionnaire. The missing results might provide more clarity.

VIII. Threats to validity

We discuss relevant threats to the validity of our experiment according to the categorization by Wohlin et al. [44].

Internal Validity: We do not regard maturation as a serious threat. With respect to tiredness and boredom, the results of the post-experiment survey show that most of the participants perceived the experiment as interesting and fun. Further, all participants contributed their posts and ratings on the Garuso platform voluntarily.

We do not perceive selection as a major threat. Due to the inherent motivation of volunteers towards the subject of an experiment, it is often believed that they might not represent the general community. However, in the context of stakeholders outside organizational reach, due to not being able to instruct them, this kind of inherent motivation is needed.

We do not consider the random assignment of the participants to the groups as a major threat. We acknowledge that the strength of inherent motivation can vary among the participants. However, motivational intensity is mainly influenced by how autonomy, competence, and relatedness are perceived during an activity or by a reward [22]. In this context, we therefore regard the gamification design as more important than the homogeneity of the groups.

External Validity: We do not regard interaction of selection and treatment as a major threat. The effects caused by game elements (and by the algorithm controlling them) depend on the context in which gamification is applied. Therefore, the results of our experiment cannot just be generalized to other application domains. However, due to the experiment design, e.g., popular game elements and activities that typically are performed on a social media platform, we think generalization is possible for most RE activities that involve stakeholders outside organizational reach for collaboration.

We regard interaction of setting and treatment as a minor threat. To be as close to reality as possible, we identified stakeholders outside organizational reach as participants, conducted a field experiment (instead of a laboratory experiment), and applied algorithms that are reasonable in the given context. However, to isolate the dependent variables we inhibited social and normative comparisons, which are regarded as catalysts in gamification. Therefore, our results are preliminary. Further, additional approaches need to be considered to deal with common challenges in RE, e.g., scalability, duplicated posts, and saturation, i.e., the decreasing number of post.

Construct Validity: We addressed mono-operation bias by running two sub-experiments in which we evaluated two independent variables with a treatment and a control group.

We do not consider mono-method bias as a threat since we evaluated the data in different ways, using descriptive statistics, observations, statistical tests and a questionnaire.

We addressed evaluation apprehension, i.e. looking better when being evaluated, by inhibiting comparisons among the participants and by assuring full confidentiality to the participants to prevent evaluation stress.

Conclusion Validity: We addressed violated assumptions of statistical tests by testing the data with respect to normal distribution and variance equality.

We addressed reliability of measures by involving senior researchers from different fields to discuss the experiment design and test the usability of the platform.

We limited the risk of false ratings by allowing participants to change their ratings at any time. Further, we randomized the order of shown posts to prevent that new posts are always shown first.

IX. Conclusion and Future Work

We report on a between subject field experiment in which we investigated the effects of gamification algorithms on requirements prioritization. Our focus was on the *effects of the algorithms* that control the game elements *points* and *levels* on the *number* and *values* of post ratings on the Garuso platform. The experiment involved 20 stakeholders outside organizational reach over a period of twelve days.

The results show that the algorithms controlling the game elements have a statistically significant influence on the stakeholders outside organizational reach with respect to their contributions to requirements prioritization. We believe that the presented research contributes important knowledge to leveraging the wisdom and creativity of a crowd of stakeholders when prioritizing requirements as well as to the body of gamification principles in the field of RE. Yet, the results are preliminary. To tap into the predicted high potential of gamification in RE [24], more research is needed. We encourage researchers to further exploit gamification algorithms with respect to other RE activities and game elements, more participants, and longer periods of time.

In a next step, we are going to apply our findings in the final implementation of the Garuso platform, conduct a real world case study, and evaluate the results with respect to stakeholder participation.

Acknowledgments

We thank all experiment participants for their voluntary engagement, our colleague Chat Wacharamanotham for advice on the experiment design, and Yolanda Schlumpf from the Department of Psychology of UZH for assistance with the statistics.

REFERENCES

[1] P. Achimugu, A. Selamat, R. Ibrahim, and M. N. Mahrin, "A systematic literature review of software requirements prioritization research," *Information and Software Technology*, vol. 56, no. 6, pp. 568–585, 2014.

[2] M. Glinz and R. J. Wieringa, "Stakeholders in Requirements Engineering," *IEEE Software*, vol. 24, no. 2, pp. 18–20, 2007.

[3] W. Maalej and D. Pagano, "On the socialness of software," in *Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing (DASC)*. IEEE, 2011, pp. 864–871.

[4] D. Yang, D. Wu, S. Koolmanojwong, A. W. Brown, and B. W. Boehm, "WikiWinWin: A wiki based system for collaborative requirements negotiation," in *41st IEEE Annual Hawaii International Conference on System Sciences*. IEEE, 2008, pp. 24–24.

[5] J. Konaté, A. E. K. Sahraoui, and G. L. Kolfschoten, "Collaborative requirements elicitation: A process-centred approach," *Group Decision and Negotiation*, vol. 23, no. 4, pp. 847–877, 2014.

[6] S. Lohmann, S. Dietzold, P. Heim, and N. Heino, "A web platform for social requirements engineering," in *Software Engineering 2009, Workshop Volume, Lecture Notes in Informatics*, vol. P-150, 2009, pp. 309–315.

[7] T. Johann and W. Maalej, "Democratic mass participation of users in requirements engineering?" in *23rd IEEE International Requirements Engineering Conference (RE)*. IEEE, 2015, pp. 256–261.

[8] J. Fernandes, D. Duarte, C. Ribeiro, C. Farinha, J. M. Pereira, and M. M. da Silva, "iThink: A game-based approach towards improving collaboration and participation in requirement elicitation," *Procedia Computer Science*, vol. 15, pp. 66–77, 2012.

[9] R. Snijders, F. Dalpiaz, S. Brinkkemper, M. Hosseini, R. Ali, and A. Ozum, "REfine: A gamified platform for participatory requirements engineering," in *1st IEEE International Workshop on Crowd-Based Requirements Engineering (CrowdRE)*. IEEE, 2015, pp. 1–6.

[10] A. Kankanhalli, M. Taher, H. Cavusoglu, and S. H. Kim, "Gamification: A new paradigm for online user engagement," in *Third International Conference on Information Systems*, 2012.

[11] K. Wiegers, "First things first: prioritizing requirements," *Software Development*, vol. 7, no. 9, pp. 48–53, 1999.

[12] P. Berander and A. Andrews, "Requirements prioritization," in *Engineering and managing software requirements*. Springer, 2005, pp. 69–94.

[13] N. Mulla and S. Girase, "A new approach to requirement elicitation based on stakeholder recommendation and collaborative filtering," *International Journal of Software Engineering & Applications*, vol. 3, no. 3, p. 51, 2012.

[14] J. Karlsson and K. Ryan, "A cost-value approach for prioritizing requirements," *IEEE Software*, vol. 14, no. 5, pp. 67–74, 1997.

[15] J. Karlsson, "Software requirements prioritizing," in *Second IEEE International Conference on Requirements Engineering*. IEEE, 1996, pp. 110–116.

[16] T. Tourwé, W. Codenie, N. Boucart, and V. Blagojević, "Demystifying release definition: from requirements prioritization to collaborative value quantification," in *International Working Conference on Requirements Engineering: Foundation for Software Quality (REFSQ)*. Springer, 2009, pp. 37–44.

[17] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: Defining gamification," in *15th International Academic MindTrek Conference: Envisioning future media environments* ACM, 2011, pp. 9–15.

[18] J. J. Lee and J. Hammer, "Gamification in education: What, how, why bother?" *Academic Exchange Quarterly*, vol. 15, no. 2, p. 146, 2011.

[19] S. Deterding, "Gamification: Designing for motivation," *Interactions* · vol. 19, no. 4, pp. 14–17, 2012.

[20] S. Kujala, M. Kauppinen, L. Lehtola, and T. Kojo, "The role of user involvement in requirements quality and project success," in *13th IEEE International Requirements Engineering Conference*. IEEE, 2005, pp. · 75–84.

[21] P. Lombriser, F. Dalpiaz, G. Lucassen, and S. Brinkkemper, "Gamified requirements engineering: Model and experimentation," in *22nd International Working Conference on Requirements Engineering: Foundation for Software Quality (REFSQ)*. Springer, 2016, pp. 171–187.

[22] R. M. Ryan and E. L. Deci, "Intrinsic and extrinsic motivations: Classic definitions and new directions," *Contemporary Educational Psychology*, vol. 25, no. 1, pp. 54–67, 2000.

[23] M. R. Lepper, D. Greene, and R. E. Nisbett, "Undermining children's intrinsic interest with extrinsic reward: A test of the overjustification hypothesis." *Journal of Personality and Social Psychology*, vol. 28, no. 1, pp. 129–137, 1973.

[24] O. Pedreira, F. García, N. Brisaboa, and M. Piattini, "Gamification in software engineering–A systematic mapping," *Information and Software Technology*, vol. 57, pp. 157–168, 2015.

[25] D. J. Dubois and G. Tamburrelli, "Understanding gamification mechanisms for software development," in *9th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering (ESEC/FSE)*. ACM, 2013, pp. 659–662.

[26] C. Ribeiro, C. Farinha, J. Pereira, and M. M. da Silva, "Gamifying requirement elicitation: Practical implications and outcomes in improving stakeholders collaboration," *Entertainment Computing*, vol. 5, no. 4, pp. 335–345, 2014.

[27] M. Z. Huber Kolpondinos and M. Glinz, "Tailoring gamification to requirements elicitation: A stakeholder centric motivation concept," in *10th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)@ICSE2017*. IEEE, 2017, pp. 9–15.

[28] J. Scott, *Social network analysis*. Sage, 2012.

[29] S. L. Lim, D. Quercia, and A. Finkelstein, "Stakesource: Harnessing the power of crowdsourcing and social networks in stakeholder analysis," in *32nd International Conference on Software Engineering – Volume 2*. ACM, 2010, pp. 239–242.

[30] B. Nuseibeh and S. Easterbrook, "Requirements engineering: a roadmap," in *The Future of Software Engineering*. ACM, 2000, pp. 35–46.

[31] M. Sailer, J. Hense, H. Mandl, and M. Klevers, "Psychological perspectives on motivation through gamification." *Journal of Interaction Design and Architecture(s) (IxD&A)*, vol. 19, pp. 28–37, 2013.

[32] Nest - exploring the future of buildings. [Online]. Available: https://www.empa.ch/web/nest/ (Accessed: 2016-08-25).

[33] V. R. Basili, "Applying the goal/question/metric paradigm in the experience factory," *Software Quality Assurance and Measurement: A Worldwide Perspective*, pp. 21–44, 1993.

[34] G. Charness, U. Gneezy, and M. A. Kuhn, "Experimental methods: Between-subject and within-subject design," *Journal of Economic Behavior & Organization*, vol. 81, no. 1, pp. 1–8, 2012.

[35] Garuso project - energy efficient workout. [Online]. Available: http://www.ifi.uzh.ch/en/rerg/research/stakeholderengagement/garuso/energyefficientworkout.html (Accessed: 2016-08-20).

[36] A. Nagle, P. Wolf, and R. Riener, "Towards a system of customized video game mechanics based on player personality: Relating the big five personality traits with difficulty adaptation in a first-person shooter game," *Entertainment Computing*, vol. 13, pp. 10–24, 2016.

[37] G. Zichermann and C. Cunningham, *Gamification by design: Implementing game mechanics in web and mobile apps*. O'Reilly Media, Inc., 2011.

[38] R. Bartle, "Virtual worlds: Why people play," *Massively Multiplayer Game Development*, vol. 2, no. 1, 2005.

[39] Stackoverflow. [Online]. Available: http://stackoverflow.com/ (Accessed: 2017-01-02).

[40] Garuso project - monitored experiment data. [Online]. Available: https://figshare.com/s/9c4a01f107cd388fd8c6

[41] F. J. Massey Jr, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[42] J. Cohen, "A power primer." *Psychological Bulletin*, vol. 112, no. 1, p. 155, 1992.

[43] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, *The measurement of meaning*. University of Illinois Press, 1964.

[44] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in software engineering*. Springer Science & Business Media, 2012.